Capítulo 2

Correlación y regresión

Si bien la correlación entre variables se deriva del análisis de regresión y, por lo tanto, éste debería exponerse primero, por razones didácticas, que podrán apreciarse en nuestra exposición, iniciaremos este capítulo con la presentación de diversos coeficientes de correlación calculables con variables proporcionales e intervales.

Coeficientes de correlación

En esta sección, vamos a presentar el principal coeficiente de correlación aplicable a variables intervales y proporcionales.

Coeficiente de correlación de Pearson

Es uno de los coeficientes más utilizados para medir la correlación que pueda existir entre dos variables de los niveles de medición que señalamos más arriba. Además del nombre de Pearson, recibe también el de momento producto y, en cuanto a la naturaleza de la covariación que mide, es un *coeficiente de orden cero* en cuanto no controla el efecto que podrían producir otras variables en la correlación. Matemáticamente se define como la razón entre la covariación y la raíz cuadrada del producto de la variación en X y la variación en Y. Consecuentemente con esta definición, su fórmula de cálculo es la siguiente:

$$r = \frac{\Sigma\left(X-\overline{X}\right)\left(Y-\overline{Y}\right)}{\sqrt{\left[\Sigma\left(X-\overline{X}\right)^2\right]\left[\Sigma\left(Y-\overline{Y}\right)^2\right]}}$$

El coeficiente puede tomar valores entre + 1 y - 1. El primero se da cuando existe correlación positiva perfecta; el segundo, cuando hay correlación negativa perfecta (a mayor valor de una de las variables, menor valor proporcional de la otra). Además de la fórmula dada más arriba suele utilizarse esta otra, en ambos casos, para datos no agrupados:

$$r = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{\left[N\sum X^2 - (\sum X)^2\right]\left[N\sum Y^2 - (\sum Y)^2\right]}}$$

Ejemplo de cálculo.

Supongamos que deseamos averiguar si existe correlación entre los años de educación de un conjunto de ocho personas y sus niveles de intolerancia social medida por una escala de actitudes. Sean estos los datos obtenidos de cada persona:

Persona	ED(X)	INT. (Y)
1	3	8
2	7	5
3	5	8
4	10	6
5	4	7
6	8	5
7	6	4
8	5	6

Con esos datos se procede a hacer los cálculo requeridos por la fórmula:

Persona	X	X^2	Y	Y^2	XY
1	3	9	8	64	24
2	7	49	5	25	35
3	5	25	8	64	40
4	10	100	6	36	60
5	4	16	7	49	28
6	8	64	5	25	40
7	6	36	4	16	24
8	5	25	6	36	30
$\sum$ 8	48	324	49	315	281

Se introducen los valores en la fórmula y con ellos se calcula el coeficiente de correlación:

$$r = \frac{(8)(281) - (48)(49)}{\sqrt{\left[(8)(324) - (48)^2\right]\left[(8)(315) - (49)^2\right]}}$$

$$r = -0{,}56$$

El valor de - 0,56 indica que existe una correlación negativa mediana entre el nivel de educación y nivel de intolerancia, es decir, las personas con mayor educación tienden a ser más intolerantes que aquellas con mayor número de años de educación.

El cuadrado del coeficiente "r" –denominado *coeficiente de determinación*– indica la variación de "Y" que es explicada por la variación de "X". En el ejemplo, $r^2 = 0{,}31$ cifra según la cual el 31% de la variación de la variable dependiente (intolerancia) es explicada por la variación de la variable independiente (educación).

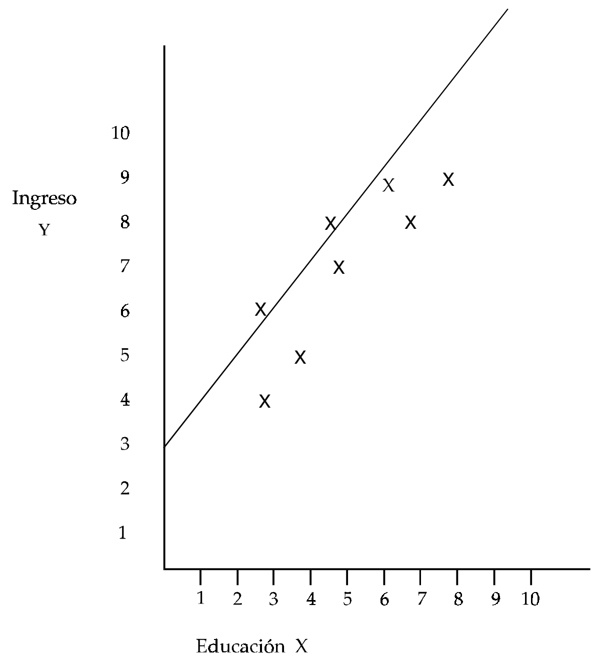
Análisis de regresión simple

Supongamos que tenemos información sobre la escolaridad ("X") y los ingresos ("Y") de ocho personas y queremos saber si existe correlación entre esas variables (proporcionales) como también saber la magnitud de los efectos de la variable independiente sobre la variable dependiente y, por fin, conocer un procedimiento para calcular el valor de esta última cuando se conocen los valores de la primera. La respuesta a la primera pregunta la obtenemos mediante la elaboración de un *diagrama de dispersión* en un eje cartesiano el cual, en el eje horizontal se representan los valores de "X" y en el eje vertical los valores de "Y". Las respuestas a las otras dos preguntas se verán a continuación de esta primera. Por el momento, supongamos, los siguientes valores de esas variables con su correspondiente diagrama de dispersión.

Escolaridad e ingreso de 8 personas

Persona	Educación: X	Ingreso: Y (en miles)
A	3	6
B	5	8
C	6	9
D	8	9
E	7	8
F	3	4
G	5	7
H	4	5

Diagramas de dispersión



En el diagrama, cada persona o caso está representado por una X que se define por el par de valores que le corresponde en cada variable. En este ejemplo puede verse que existe correlación positiva entre educación e ingreso ya que los valores de ambas varían en un mismo sentido (los valores de "Y" suben cuando suben los correspondientes valores

de "X"). De manera general, la correlación se puede expresar mediante una función matemática, que aquí recibe el nombre de *ecuación de regresión*, con la siguiente fórmula:

$$Y = a + bx$$

Cuando la relación entre dos variables es total, a cada valor de "X" le corresponde un solo valor de "Y". Sabemos, sin embargo, que en la realidad social se está muy lejos de darse esa situación. Así, para volver a la variable de nuestro ejemplo, no todas las personas con un mismo nivel educativo tienen el mismo salario, sino que todos oscilan alrededor del medio aritmético de ellos. Debido a esa variación, la ecuación de regresión que representa datos reales contiene un llamado *término de error* que suele indicarse con la letra "e":

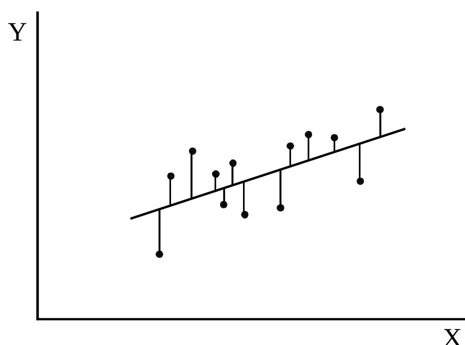
$$Y = a + bx + e$$

En la ecuación el término "a" se denomina *intercepto* e indica el punto en el cual la recta de regresión corta el eje de las "Y". En tal punto el valor de "X" es cero. El término "b" se denomina *pendiente de la recta o coeficiente b* que indica, por su lado, el número de unidades de cambio que se produce en "Y" cada vez que "X" varía en una unidad. Cuando "b" es negativo, la pendiente es decreciente ya que al aumentar el valor de "X" disminuye el correspondiente valor de "Y", y viceversa. En otras palabras, la correlación entre ambas variables es negativa.

Criterio de los cuadrados mínimos

Cuando se tiene un conjunto de puntos en un diagrama de dispersión es posible hacer pasar por ellos varias rectas. Sin embargo, debemos elegir aquella que mejor se ajusta a los datos. Para hacer esta elección se utiliza el criterio general denominado de los cuadrados mínimos según el cual la suma de los cuadrados de las desviaciones de los valores reales de "Y" a la línea es mínima. Según esto, si se trazan líneas verticales de cada uno de los puntos a la línea de los cuadrados mínimos y si se elevan al cuadrado esas distancias y se las suma, la suma resultante será menor que otra suma de cuadrados posible de trazar.

Para trazar la línea de los cuadrados mínimos, se necesita computar los coeficientes "a" y "b" de la ecuación de regresión, valores que determinan la línea con la propiedad ya



señalada. Este problema se resuelve por medio del cálculo (mediante derivadas parciales de “a” y “b”) que da las siguientes dos *ecuaciones normales:*

$$a = \bar{y} - b\bar{x}$$

$$b = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$$

El numerador de “b” ($\sum(x-\bar{x})(y-\bar{y})$) se denomina la *covariación* de X e Y: indica cómo esas variables varían en conjunto. Si se divide esa expresión por N se obtiene la llamada *covarianza*. Según esto, se puede ver fácilmente que el coeficiente “b” es igual a la razón entre la covarianza de X e Y y la varianza de X.

Sin perjuicio a lo afirmado más arriba los coeficientes “a” y “b” se pueden calcular también con estas fórmulas:

$$a = \frac{\sum Y - b\ (\sum X)}{N}$$

$$b = \frac{N\sum XY-(\sum X)(\sum Y)}{N\sum X^2 - (\sum X)^2}$$

Ejemplo de cálculo

Un diagrama de dispersión ha mostrado que hay una relación lineal entre porcentajes de trabajadores industriales en la población económicamente activa de 10 municipios y los correspondientes porcentajes de votación por el partido A, de orientación progresista. Se debe saber: a) Cuál es la variación de la variable dependiente “votación” por cada unidad

de variación de la variable independiente “porcentaje de trabajadores industriales”. b) Qué valores toma la variable dependiente cuando se toman otros valores de la variable “porcentaje de trabajadores industriales” no incluidos en los datos proporcionados (predicción de valores de y).

Comunidad	Trabajadores industriales (x) %	Votación (y) %	y^2	x^2	xy
A	12	15	225	144	180
B	6	8	64	36	48
C	8	12	144	64	96
D	4	9	81	16	36
E	5	6	36	25	30
F	6	10	100	36	60
G	5	9	81	25	45
H	10	14	196	100	140
I	4	10	100	16	40
J	7	8	64	49	56
Σ	67	101	1.091	511	751

El reemplazo de los valores del cuadro, para el cálculo de los coeficientes, da los siguientes resultados:
La ecuación de regresión es:

$$y' = 4,2 + 0,88x$$

$$b = \frac{10\,(751) - (67)(101)}{10(511) - (67)^2}$$
$$= 0,88$$
$$a = \frac{(101) - (0.88)(67)}{10}$$
$$= 4,2$$

Según esta ecuación, las respuestas a nuestro problema son: a) Cuando el porcentaje de trabajadores industriales varía en una unidad, el porcentaje de votación por el partido

A aumenta 0,88 de la unidad. b) Para un valor 3 de la variable trabajadores industriales, podemos predecir que el porcentaje de votación será:

$$
\begin{aligned}
y' &= 4{,}2 + 0{,}88\ (3)\\
&= 6{,}8\%
\end{aligned}
$$

Supuestos del análisis de regresión

El análisis de regresión trabaja con ciertos supuestos o condiciones de las cuales se destacan las siguientes:

1. Las variables independientes y las variables dependientes deben estar medidas en escalas intervales o de razón (variables cuantitativas continuas).
2. Los valores de la variable dependiente deben estar normalmente distribuidos en la población de la cual proceden.
3. Los valores del término de error tienen una distribución normal.
4. Los valores del término de error no están correlacionados con los valores de la variable independiente.
5. Las desviaciones estándar de los valores de la variable dependiente que corresponden a cada valor de la variable independiente son iguales entre sí. Esta condición recibe el nombre de supuesto de *homocedasticidad*.

Cuando se cumplen los supuestos anteriores es posible obtener buenos estimadores de los parámetros “a” y “b” en la población a partir de sus correspondientes valores en la muestra. En particular, el supuesto de la distribución normal del término de error apoya la posibilidad de efectuar pruebas de significación con relación a los parámetros mencionados. Por su lado, el supuesto de homocedasticidad también se relaciona con las pruebas de significación: su violación puede significar que tales pruebas no sean válidas. Si se tuviera constancia que existe esa violación, es decir, que se produce una condición de *heterocedasticidad*, debería utilizarse un procedimiento de ajuste de *menores cuadrados ponderados*.

Con toda la importancia que se debe dar a los supuestos señalados, se ha demostrado empíricamente que las estadísticas “F” y “t” que se utilizan en las pruebas de significación dentro del análisis de regresión son suficientemente fuertes para soportar violaciones a algunos de ellos.

Bibliografía recomendada para la unidad

Blalock, Hubert. Varias ediciones en español desde l966. *Estadística social*. Capítulo XV: Escalas nominales: Problemas de contingencia (hasta p. 259).

Briones, Guillermo. *Métodos y técnicas avanzadas de investigación aplicadas a la educación y a las ciencias sociales*. *Módulo 4: Análisis e interpretación de datos cuantitativos*. Bogotá: ICFES-PIIE, 1989. Cap. IV: Análisis de relaciones simples.

Briones, Guillermo. (Varias impresiones). *Métodos y técnicas de investigación para las ciencias sociales*. México: Trillas, 1992. Cap. 18: Medidas de asociación para variables nominales y ordinales; Cap. 19: Regresión y correlación.

Hernández Sampieri, Roberto y otros. *Metodología de la investigación*. México: McGraw-Hill, 1991. Cap. 10, sección 10.8: Análisis no paramétrico.

Loether, Herman y McTavish Donald G. *Descriptive Statistics for Sociologists*, Boston: Allyn and Bacon, 1979. Cap. 7: Measures of Association for Nominal, Ordinal, and Interval Variables.

Sierra Bravo, Restituto. *Técnicas de investigación social. Teoría y ejercicios*. Madrid: Paraninfo, 1985. Cap. 24: Asociación entre variables.